Formal Analysis and Estimation of Chance in Datasets Based on Their Properties

Journal Item

Formal Analysis and Estimation of Chance in Datasets Based on Their Properties

Abdel Aziz Taha, Luca Pappariello, Alexandros Bampoulidis, Petr Knoth, Mihai Lupu

Abstract—Machine learning research, particularly in genomics, is often based on wide shaped datasets, i.e. datasets having a large number of features, but a small number of samples. Such configurations raise the possibility of chance influence (the increase of measured accuracy due to chance correlations) on the learning process and the evaluation results. Prior research underlined the problem of generalization of models obtained based on such data. In this paper, we investigate the influence of chance on prediction and show its significant effects on wide shaped datasets. First, we empirically demonstrate how significant the influence of chance in such datasets is by showing that prediction models trained on thousands of randomly generated datasets can achieve high accuracy. This is the case even when using cross-validation. We then provide a formal analysis of chance influence and design formal chance influence estimators based on the dataset parameters, namely its sample size, the number of features, the number of classes and the class distribution. Finally, we provide an in-depth discussion of the formal analysis including applications of the findings and recommendations on chance influence mitigation.

Index Terms—High-dimensional data, Chance Correlation, Formal Estimation of chance, Generalization, Sparse Data, Genomics

1 Introduction

Datasets with a large number of variables, but at the same time a small number of samples, are being frequently used in applications of machine learning (ML) techniques in the medical domain. One example is the analysis of genomic data, where datasets typically consist of thousands of genes. One of the main problems with this type of data is that the results obtained from its analysis are hardly reproducible. For instance, subsets of features reported by one research group as predictive of some disease either largely differ from other groups' results or are not predictive when applied on other groups' data [1], [2], [3].

Michiels et. al [4] provide an emblematic example, where seven studies that claim to predict cancer based on microarray data have been reanalyzed. They reported that the results of most of these studies are overoptimistic and five of them provide prediction methods that are not better than a random predictor. In particular, they reported instability in the feature selection in the sense that the features selected by the algorithms as predictive regarding the underlying outcomes significantly change depending on the patients considered in the training sets, such that the feature selection can be described as unrepeatable. They also reported that this instability decreases with increasing the number of samples used for training and evaluation. Hua et. al [5] emphasized the considerable influence of the ratio between the number of features and the sample size on the reliability and reproducibility of results.

Lian et. al [6] observe that much research has been done on dimensionality reduction to overcome difficulties stemming from high-dimensional data, like efficiency, curse of dimensionality, and loss of performance, but less research addresses the evaluation of machine learning methods when applied to high-dimensional data. Kim et. al [7] provide an analysis showing how unstable feature selection methods are when applied to high-dimensional data. They also provide a statistical evaluation measure that incorporates the stability of selected feature subsets.

The situation of datasets consisting of a large number of features and a small number of samples is a critical situation combining both high dimensionality and low number of instances. This situation creates a kind of data sparsity, which is directly related to the high dimensionality scenario. The difficulties related to this situation stem from the curse of dimensionality [8], [9], [10], [11], [12], which will be discussed in more detail in the related work.

Such a setting (larger number of features, small number of instances) is most common in the biomedical domain [13], [14], [15]. The direct motivation for the research presented in this paper stems from our key observation while performing experiments on an RNA genomic dataset containing microarray gene expressions of 80 patients who died as a result of neuroblastoma cancer after different survival times. Our task was to predict the survival time (from diagnosis until death) based solely on the RNA data, which consists of about 16,000 features (gene expressions). Using a simple regression model in a cross-validation setup after performing a feature selection, we were able to achieve a prediction accuracy of more than 97%. Since we were doubtful of this high accuracy given the small number of samples, we questioned our result and investigated its origins in the following way. We replaced all the gene data with random numbers uniformly distributed in $[0, 1]$ and kept the target (survival time) unchanged. After applying exactly the same feature selection and regression algorithms on the random dataset, we were still surprisingly able to predict the survival time with an accuracy of about 95%. This empirical observation of the models being trained on

• *All authors are with Research Studio Data Science, RSA FG, Vienna, Austria; Petr Knoth is also with KMi, The Open University, Milton Keynes; UK E-mail: abdel.taha@researchstudio.at*

random data and even so predicting with high accuracy is a clear signal that there is something going wrong with the evaluation process (in this case cross-validation) under these experiment settings.

Combining this observation with abundant literature on this issue (e.g. [1], [4], [5], [7], [16], [17], [18], [19], [20]) reveals the need to standardize the learning/evaluation process when dealing with such extremely shaped datasets to ensure comparability of results and a clear definition of the baselines. There is also a strong evidence to directly relate this generalizability to the influence of chance correlation and thus to highlight the need for chance estimation methods and guidelines for avoiding/mitigating this influence to quantify a clear baseline and comparability.

We define chance influence as an increase in the performance of a model stemming from chance correlation in the underlying data. Our work is based on the assumption that the amount of information contained in a truly randomly generated dataset with randomly generated target classes is extremely low and can be neglected. This assumption can be justified using the basics of the information theory, namely the mutual information [21] between the features and the target—this is indeed zero in the above mentioned setup. To measure the chance influence in a dataset with particular dimensions, a model is trained to predict the target class on a randomly generated dataset of these dimensions by only exploiting chance correlations. The accuracy achieved by such model is assumed to be a measure of the chance influence on a (real) dataset of these dimensions. One of the main goals of this document is to provide a formal estimator of the chance influence in a dataset, which we will call ϕ (phi).

In this paper, we first empirically demonstrate the enormous impact of chance correlation on training and evaluation of ML algorithms in high-dimensional detests with low numbers of examples. We show that cross-validation procedures do not remove the possibility of obtaining highly overoptimistic results as the chance correlation phenomenon is not related to over-fitting. Our observation is confirmed by running thousands of simulated experiments on random datasets of different dimensionality and across different data types and classification/regression tasks. We also provide a formal analysis of chance and mathematically model the relation between chance extent and the properties of the dataset (i.e. number of features, number of samples, and number of classes), which results in formal chance estimators. Finally, we provide a discussion of chance correlation including feature selection under extreme settings and guidelines to mitigate the influence of chance on prediction models [22].

2 Related Work and background

2.1 Curse of Dimensionality

The curse of dimensionality [9], [12] refers to the problems arising when dealing with high dimensional data, namely the fact that optimally estimating a function would require parsing the complete data space and, thus, requiring an unrealistic amount of computational resources [12]. The curse of dimensionality has had an impact on many research fields and has motivated a lot of research that mitigates its effects. However, the curse of dimensionality has several aspects, two of which we are describing in this section.

One aspect of the curse of dimensionality is distance concentration. This is the phenomenon that could occur in high dimensional data spaces, with which distances between points become too similar, thereby reducing the utility of the information contained in the data space. Distance concentration negatively affects tasks that rely on distance and similarity measures, and considerable research has been conducted to mitigate its effects in these tasks [9], [23], [24], [25], [26], [27], [28]. In the context of unsupervised machine learning, dimensionality reduction methods, such as Principal Component Analysis (PCA) [29] and Factor Analysis (FA) [30], are typically used to mitigate the effects of this phenomenon.

Another aspect of the curse of dimensionality in supervised learning is data sparsity: the higher the number of features, the higher the sample size is required, in order for the ML model to be generalisable. Similarly to unsupervised learning, data sparsity in supervised learning is mitigated by dimensionality reduction methods [31], [32], [33], many of which are typically found in popular software libraries [34], [35], [36]. Our work analyzes the difficulties arising from settings that promote data sparsity, namely when training ML models based on data that combines both high dimensionality and low number of instances.

2.2 Learning Curves

Learning curves refer to a class of approaches that have been used to tackle problems with supervised learning related to the dimensionality of the dataset used, i.e. the ratio between the number of features and the number of instances.For this, one commonly employs learning curves to predict how the classification accuracy would change when the sample size is increased.

A learning curve is a model that describes the progress of a learning process, e.g. the accuracy of a ML algorithm as a function of the number of examples used in the learning phase. A common method to implement a learning curve is to fit an inverse power law curve using a small number of samples [20], i.e.:

$$e(n) = an^{-\alpha} + b, \tag{1}$$

where e is the error rate given n training samples, a is the learning rate, b the Bayes error, and α the decay rate. The values of these parameters depend on the classification algorithm and the dataset.

Many approaches follow this principle to predict the accuracy of an algorithm in a confidence interval around the learning curve given the number of samples, e.g. Figueroa et. al [37]. Others estimate the minimum number of samples required for a classifier to keep the error in a particular confidence interval, e.g. Mukherjee et. al [20], Dobbin et. al [19], and Beiletes et al. [38]. The loss function is another way often used to monitor the learning progress to get feedback about, among others, the model quality achieved given the data.

However, all these approaches aim at optimizing the accuracy by finding the optimal number of samples. But optimizing the accuracy does not necessarily improve the

model generalizability, especially if a part of it is the result of chance correlations. It is even the opposite: We show in this work that increasing the number of features while keeping the number of rows constant most likely leads to an increase in accuracy by chance.

2.3 Bias Caused by Feature Selection

Ambroise et. al [17] thoroughly discusses feature selection bias performed prior to cross-validation, when such feature selection is performed on the entire dataset. They state that in this case the estimation of the prediction error is too optimistic. This is because the testing is influenced by the selection bias stemming from the fact that the test set is a subset of the entire set used for feature selection. As bias correction, they suggest using a special cross-validation and bootstrapping method.

Ein-Dor et. al [1] investigated the common problem of robustness in feature selection procedures in genomics research, i.e. the problem of the gene subsets identified as predictive of an outcome being not stable and depending on the samples included in the training. These subsets of genes identified using different training samples are not only different, but even the overlap (common genes) is very small. Furthermore these gene lists are in general significantly less predictive when applied on external datasets. They reported that the problem of robustness can be mitigated by using more samples in the training. The authors provided a formal method to find the number of samples required to achieve a particular overlap between gene lists identified based on different training sets. This model was achieved based on the assumption that the overlap is a random variable with a normal distribution. Their results show that thousands of samples are required to achieve an overlap of more than 50% between the gene lists.

While this research denotes the importance of how the feature selection is performed to mitigate chance correlation and Kalousis et al. [39] even propose a measurement of the stability of feature selection algorithms, they do not provide methods to estimate and quantify the chance given the parameters of a dataset, such as the number of features, the sample size, and the number of classes.

Clark et. al [40] empirically investigated the influence of chance correlation on partial least square (PLS). They demonstrated how the chance influence increases with increasing the number of features. Kuligowski et. al [16] investigated the prediction accuracy in metabolomics using partial least squares discriminant analysis. They reported that cross-validation after feature selection provides overoptimistic results due to chance correlation. The effect of chance correlation is expressed by means of p-values calculated by using a permutation test that include the variable selection. Taha et al. [41] (the precursor to this paper) show that the influence of chance is considerable in datasets with a large number of features and it can lead to non-generalisable models. They show that it depends on the way feature selection is performed and observe that the influence of chance decreases when the number of classes increases.

2.4 Critical Research

In recent years, attention has been drawn to the issue of irreproducibility of studies across several areas of science [3], [42], [43], [44], [45], [46]. Ioannidis [3] has indeed reported that many published results, most notably biomedical papers, cannot be reproduced by other researchers. In many cases, a widely spread misinterpretation of p-values lies at the heart of this problem. This led a number of authors to declare a discovery where, in fact, only a random chance was observed [45], [46]. Colquhoun [45], [46] showed that the common practice of assessing *statistical significance* based on p-values (typically $p \leq 0.05$), which is often wrongly thought of as the probability of the result occurring by chance, generally leads to false positive rates higher than 5%.

3 Notation

We provide definitions of ML processes based on randomly generated datasets, i.e. datasets where the features (variables) are meaningless with respect to the target class. Analysing many random datasets with varying parameters enables us to observe and measure the effect of chance correlation.

Definition 1. *Classification from random*: Let $D = \{F_1, \ldots, F_m, F^*\}$ be a random dataset of the shape $n \times (m+1)$ where F_1 to F_m are columns (features) in the dataset (we will refer to them, in short, as F) and F^* is the target class column that partitions all n instances into r classes $q_1, \ldots, q_r$ of sizes $Q_1, \ldots, Q_r$. The categorical values of the features and the r classes are generated and assigned to the target randomly. Classification models can be trained on this dataset to predict the target classes F^*.

Definition 2. *Regression from random*: Let D be a random dataset like in Definition 1, except that F^* is a numeric target value. Regression models can be trained on this random dataset to predict the numeric target value.

Definition 3. *Learning from shuffled real dataset*: Let $D^S = \{F_1, ..., F_m, F^S\}$ be a dataset whose shape is like the one described in Definition 1, but containing real data, i.e. F_1 to F_m are *not* randomly generated. D^S is modified by replacing F^* with F^S, where F^S is obtained by randomly shuffling the vector F^*. As a result, D^S becomes meaningless regarding F^*.

In this paper, the *shape* of a nominal dataset is given by three parameters, namely the number of rows (samples) n, the number of columns (features m), and the number of classes r, i.e. the cardinality of a set of unique values in the F^* . The shape $n \times m \times r$ denotes a dataset consisting of n rows (each row referring to a data sample), m columns (each column referring to a feature, target excluded), and r classes, where each sample belongs to one class. We define $\rho = m/n$ as the ratio of the number of features and the number of samples. Furthermore, we use the term *wide dataset* to denote a dataset with $\rho > 10$.

4 Chance influence on Prediction

In this section, we show that the prediction accuracy of models trained using wide datasets can be, to a large extent, influenced by chance. First, we empirically demonstrate this claim by training a large number of algorithms using

random datasets and showing their accuracies. Second, we provide a formal expression that measures the extent of chance as a function of the dataset parameters (shape), as well as approximated estimators of chance that can easily and efficiently be computed. We do our analysis empirically and theoretically in parallel as follows:

I. In the empirical part, we generate random datasets according to Definition 1. We train classification models, evaluate their performances, and analyze the results in relation to the dataset parameters, namely its shape and class distribution.
II. In the theoretical part, we provide a formal analysis of chance in datasets in the form of formal expression of chance estimators as a function of the dataset parameters.

4.1 Demonstration of chance impact on prediction

For demonstrating the accuracy obtained by chance, we generated a group of random datasets (RDCAT) to be used for the empirical analysis of classification. These are 1,000 datasets generated according to Definition 1. Each dataset has a sample size varying from 10 to 1000 samples, i.e. $n \in [0, 1000]$. The number of features in each dataset is also varying in the same interval, i.e. $m \in [0, 1000]$. The number of classes r varies from 2 to 9. To achieve class imbalance, the size Q_i of each class q_i is selected randomly from 1 to $n - r + 1$ under the condition that $\sum_{i=1}^{r} Q_i = n$.

Using the random datasets of RDCAT, we trained 1,000 classification models. We use Best First Search combined with an Information Gain evaluator as the feature selection method and a J48 Tree as the classification algorithm in a stratified 10-fold cross-validation evaluation process using the Weka framework[1]. The predicted classes are evaluated against the true target classes to find the classification performance obtained purely by chance. This evaluation is done using the F-measure, which is a standard (overlap) metric for evaluating class-imbalanced problems.

Figure 1 shows the accuracies of 1,000 J48 classification models (white diamonds), where each of them has been trained on one dataset from the RDCAT group. The experiments are sorted first according to the number of classes r and then according to model accuracy within each class number. The first interesting observation is the high model accuracies obtained from learning from random data. These accuracies are achieved despite applying 10-fold cross-validation, which is a standard method of evaluation. Second, the figure shows a strong negative relation between prediction accuracy and the number of classes r. The more classes, the lower the achieved accuracy. Third, within each plot for fixed r there is a strong variation in accuracy. We will show in the next sections that this variation depends on the other shape parameters, namely ρ (the ratio between features and sample numbers) and class imbalance. Fourth, the impact of these parameters decays with increasing r, which is reflected by the decrease in variance of the accuracy measurements with increasing r (cf. Fig. 1).

1. https://www.cs.waikato.ac.nz/ml/weka/

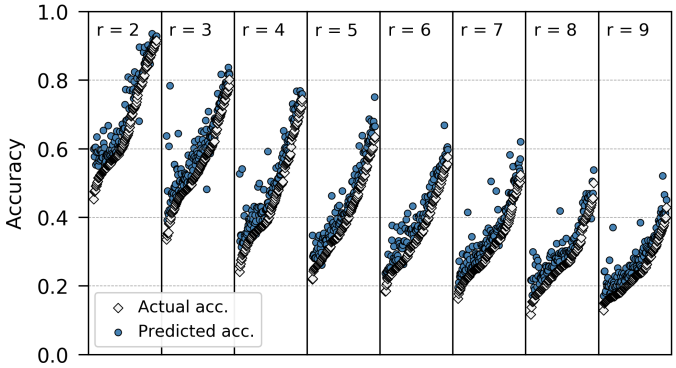

Fig. 1. Classification performance (white diamonds) in terms of F-score and predicted accuracy (blue circles) for models trained on the 1,000 random datasets of the RDCAT group. Shown is the classification performance for different number of classes r (from 2 to 9). The results are sorted first according to r and then according to model accuracy within each class number.

4.2 Formal estimation of chance on classification

Given this empirical observations in Sec.4.1, can we estimate these scores without performing the actual simulations? The aim of this section is to formally calculate the influence of chance on classification models based on the dataset parameters (dataset shape and class distribution), namely to estimate the prediction accuracy of a classification model by chance, i.e. when trained on a randomly generated dataset.

We consider the expected correlation by chance between the feature values and the target class values. Since we are talking about a classification task, which means nominal class values, we consider the probability of match (i.e. overlap) by chance between class values and feature values as a measure of correlation.

To this end, consider a dataset according to Definition 1 that consists of m features and n samples, which are divided into r classes $q_1, \ldots, q_r$ of size $Q_1, \ldots, Q_r$, i.e. $\sum_{i=1}^{r} Q_i = n$. In order to simplify the computations, we start with rearranging the rows of the dataset D, i.e. both F and F^*, by grouping the ones belonging to the same class in F^*. The target vector will thus have the following form:

$$F^* = [\underbrace{q_1, \ldots, q_1}_{Q_1}; \ldots; \underbrace{q_r, \ldots, q_r}_{Q_r}]^\top.$$

First, consider a single block Q_i of a single column of F. We then define the random variable (RV) X_j^i to take the value 1 if the j-th row in Q_i predicts the right class, that is, if it matches the corresponding element of F^*, and 0 otherwise. This is modelled by a Bernoulli trial with success probability $p_i = Q_i/n$, i.e. $X_j^i \sim \text{Ber}(p_i)$. Extending the same reasoning to the whole block Q_i, we see that

$$X^i = \sum_{j=1}^{Q_i} X_j^i \sim \text{B}(Q_i, p_i),$$

where $B(\cdot, \cdot)$ refers to the Binomial distribution. The RV X^i gives the number of rows of the sector Q_i with the right category. The total number of rows having the right category is obtained by putting together all the blocks, i.e. $X = \sum_{i=1}^{r} X^i$, where each element $X^i \sim \text{B}(Q_i, p_i)$. Note, however, that X is *not* simply distributed according to a

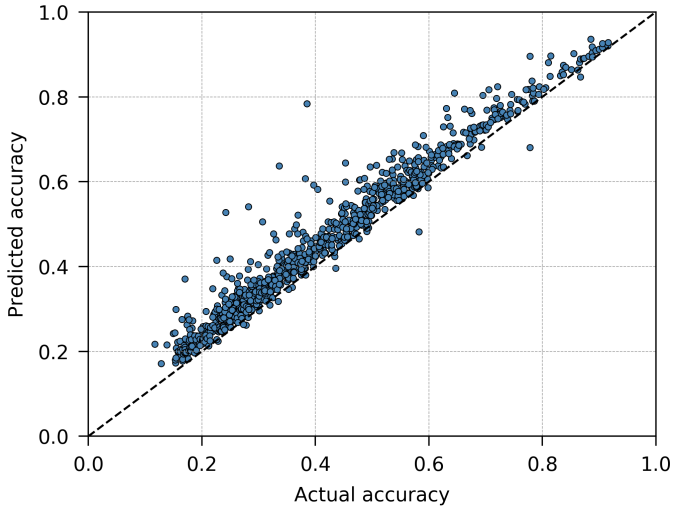

Fig. 2. Q-Q plot showing the predicted accuracy against the actual one for 1000 RDCAT group datasets. Shown are the results of the exact formula [Eq. (4)], which reveals a Pearson correlation coefficient of 0.979 (95% CI $[0.976, 0.981]$) with the experimental results.

Binomial distribution. Its distribution is instead given by the following convolution of the individual distributions:

$$p(X=k) = p\Big(\sum_{i=1}^{r} X^i = k\Big) = \sum_{\substack{(i_1,\ldots,i_r)\in\mathbb{N}^r \\ i_1+\ldots+i_r=k}} \prod_{j=1}^{r} \binom{Q_j}{i_j} p_j^{i_j}(1-p_j)^{Q_j-i_j}, \quad (2)$$

where $p_j = Q_j/n$.

We generalize now the above discussion by considering all the m columns of the dataset. More precisely, we consider the random variable $\overline{X}$ to take the value 1 if at least one of the columns in D correctly predicts at least k entries of F^*, and 0 otherwise. Using the fact that $p(\overline{X}=1) = 1 - p(\overline{X}=0)$, together with $p(\overline{X}=0) = p(X<k)^m$ and Eq. (2), implies that $\overline{X} \sim \text{Ber}(\bar{p}_k)$, where

$$\bar{p}_k = 1 - \Bigg[\sum_{\substack{(i_1,\ldots,i_r)\in\mathbb{N}^r \\ i_1+\ldots+i_r<k}} \prod_{j=1}^{r} \binom{Q_j}{i_j} p_j^{i_j}(1-p_j)^{Q_j-i_j}\Bigg]^m. \quad (3)$$

We then define the chance estimator of a classification model, which is denoted by ϕ, as the expectation value of the RV $\overline{X}$ averaged over the number of instances, i.e. by

$$\phi_c = \frac{1}{n}\sum_{k=1}^{n} \bar{p}_k\,, \quad (4)$$

with $\bar{p}_k$ given by Eq. (3). Here we have used the fact that the expectation value of $\overline{X}$ is $E[\overline{X}] = \bar{p}_k$. We will use Eq. (4) to predict the results of all the experiments with the random datasets.

To validate the performance of the chance estimator [i.e. Eq. (4)], we applied it on all 1000 datasets of the RDCAT group and compared them with the actual accuracies of the corresponding learning models. Figure 1 shows the predicted accuracies obtained from Eq. (4) and the actual accuracies as obtained from the experiments. The quality of our predictions is further confirmed by Figure 2, which shows the Q-Q plot of the correlation between the actual and predicted accuracies and reveals a Pearson correlation coefficient of 0.979 (95% CI $[0.976, 0.981]$).

Eq. (4) is a summation of elements, each of which is a summation over all possible partitions of k [see Eq. (3)], which is rather involved, especially when efficiency is a key factor. Therefore we introduce estimators that are easier to calculate, less precise, but often sufficient in the practise. From Eq. (3), we can see that part of the complexity stems from the individual probabilities p_j, i.e. considering an individual success probability for each class q_j (class imbalance). Therefore, our simplification strategy is to replace p_j with one single representative probability. In this case Eq. (3) reduces to

$$\tilde{p}_k = 1 - \Bigg[\sum_{t=0}^{k-1} \binom{n}{t} p^t (1-p)^{n-t}\Bigg]^m, \quad (5)$$

where p is the single representative property.

Now based on Eq. (5), we define two estimators by choosing representative probabilities p. The first estimator, which we denote as the *pessimistic* estimator $\breve{\phi}$, serves as a sort of lower bound of the exact estimator in Eq. (4). The second one, which we denote the *optimistic* estimator $\hat{\phi}$, serves instead as an upper bound.

For the pessimistic estimator, we consider $p = 1/r$ as the representative probability, i.e. we assume that there is no class imbalance in the dataset. This corresponds to the case in which $Q_j = n/r\ \forall j$, which means that $p_j = 1/r\ \forall j$. This results in [note that we directly substitute Eq. (5) into Eq. (4)]:

$$\breve{\phi} = 1 - \frac{1}{n}\sum_{k=1}^{n}\Bigg[\sum_{t=0}^{k-1}\binom{n}{t} p_{\text{min}}^t (1-p_{\text{min}})^{n-t}\Bigg]^m, \quad (6)$$

where $p_{\text{min}} = \frac{1}{r}$. Since class imbalance is a crucial factor for determining the influence of chance correlation, the assumption of zero class imbalance leads to a severe underestimation of chance effects.

For the optimistic estimator, we consider the maximum probability as a representative probability, i.e. the probability corresponding to the class with the largest size, which is given by $p = \max(Q_j)/n$. This results in

$$\hat{\phi} = 1 - \frac{1}{n}\sum_{k=1}^{n}\Bigg[\sum_{t=0}^{k-1}\binom{n}{t} p_{\text{max}}^t (1-p_{\text{max}})^{n-t}\Bigg]^m, \quad (7)$$

where $p_{\text{max}} = \max(Q_j)/n$. $\hat{\phi}$ overestimates the impact of chance, as the one that mostly weights the class imbalance.

Figure 3 shows the Q-Q plot of the results of both the pessimistic [Eq. (6)] and optimistic [Eq. (7)] estimators, together with the original estimator [i.e. Eq. (3), after plugging it into Eq. (4)]. We see that these two estimators—the pessimistic shown by orange, down-pointing triangles and the optimistic represented by green, up-pointing triangles—envelope the original one.

5 Discussion

In this section, we analyze and discuss the methods presented in Sec. 4 from different viewpoints to provide concrete results and conclusions. The first three sections empirically show the impact of the different factors on the chance

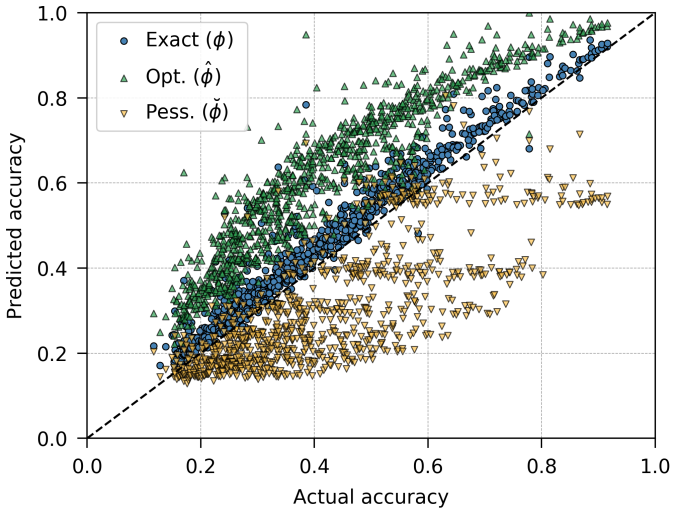

Fig. 3. Q-Q plot showing the predicted accuracy against the actual one for the 1000 datasets of the RDCAT datasets. Shown are the results of the exact formula ϕ_c (blue circles) and its upper and lower bounds (green, up-pointing and orange, down-pointing triangles), i.e. the optimistic ($\hat{\phi}$), and the pessimistic ($\check{\phi}$) estimators. ϕ_c, $\hat{\phi}$ and $\check{\phi}$ reveal Pearson correlation coefficients with the measured accuracy of 0.979 (95% CI $[0.976, 0.981]$), 0.931 (95% CI $[0.922, 0.939]$), and 0.753 (95% CI $[0.724, 0.779]$), respectively.

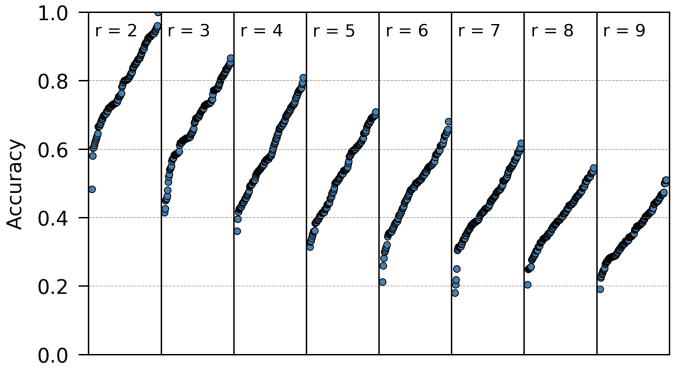

Fig. 4. Classification performance in terms of F-score for models trained on the 1,000 datasets of the RDCAT group. The data points are first sorted according to r and then according to their score. The number of classes r varies from 2 to 9. It can be seen that this parameter defines clear levels of accuracy.

in a dataset. Sec. 5.1 demonstrates the impact of the number of classes r on the extent of chance, Sec. 5.2 investigates the impact of class size and class imbalance, while Sec. 5.3 reveals the impact of the columns-rows ratio ρ. In Sec. 5.4, we test the estimators using modified real data instead of purely synthetic data to rule out the possibility that the performance of the estimators is subject to particular synthetic distributions. In Sec. 5.5, we motivate extending the analysis to regression in future work by showing that the accuracy of regression models trained on random numeric data can be estimated using the proposed estimators, although they were designed for classification. In Sec. 5.6, we suggest the use of the well-known normal approximation to reduce the computational complexity of the proposed estimators. In Sec. 5.7, we show how the proposed estimators can be used to correct for chance. We take a deeper look at the Kappa measure and show that it cannot correct for the chance stemming from high dimensionality. In Sec. 5.8, we provide guidelines to mitigate chance influence in wide datasets based on the findings of this analysis. Finally, in Sec. 5.9, we summarize value and impact areas of this analysis.

5.1 Impact of number of classes r

The accuracies of classification models trained on random datasets with different numbers of classes r is shown in Fig. 4. Regardless of the other parameters governing the extent of chance, r defines clear levels of accuracy. In each level, there is a basis accuracy of $1/r$, which is where the accuracy range starts, and an additional accuracy caused by the other factors (dimensionality and imbalance). Interestingly, this additional contribution is suppressed by the level (r), i.e. the accuracy range gets narrower with increasing r, which means that the more classes there are, the less additional accuracy occurs. This observation leads to the important conclusion that classification models with many classes are significantly more resilient to chance.

5.2 Impact of class size and class imbalance

In this section, we empirically show that class size and class imbalance have a considerable impact on the amount of chance in a dataset. For this, we consider the F-measure for specific classes rather than the weighted F-measure over all classes, as was the case in the previous sections. To this end, we trained models on the random datasets in the RDCAT group, thereby recording three values for each model: (i) The F-measure of each class, (ii) the size of the class as a percentage of the dataset size, and (iii) the number of classes in the dataset.

Figure 5 shows the the accuracies in terms of F-measure for the individual classes, where the latter are sorted according to their percentage size, i.e. Q/n, where Q is the class size and n is the number of samples. Each point corresponds to a class in a dataset. The x-axis shows the size (in percentage) of the class and the number of classes in the corresponding dataset. The y-axis shows the F-measure for that class. The figure clearly shows the large impact of the class size on the accuracy by chance: classes with a relative size smaller than 10% have very low accuracies, while those with larger sizes have significantly higher accuracies reaching up to nearly 1. This is expected and intuitive, since a naive classifier that always predicts the significantly larger class will perform very well.

It is noticeable that the spread of the scores achieved on classes of intermediate size (see central region in Fig. 5) is much larger than that obtained on small (left region) and large (right region) classes. A given, intermediate class size (i.e. a vertical cut in Fig. 5) gives thus rise to a wide range of accuracies. For instance, at a fixed class size of 30% there is a spread of F-scores across a range of more than 0.4. This variety obviously stems from factors in the data set other than class size and number of classes.

To investigate where the variation of accuracy for fixed class size is stemming from, we first sorted the values according to class size and then to number of classes in the corresponding dataset. Figure 6 (A) shows the same data as

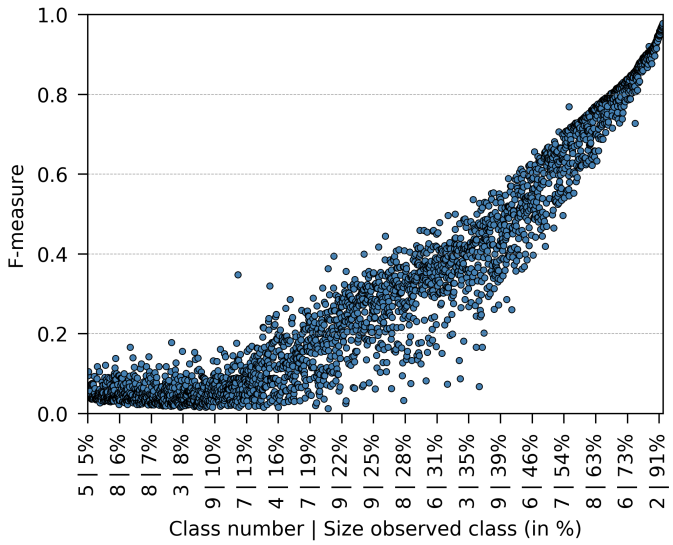

Fig. 5. Relation between accuracy and class size. The F-measure of the individual classes is depicted, rather than the weighted F-measure over all classes. On the x-axis, experiments are sorted according to their class size. For each measurement, the total number of classes in the dataset as well as the size of the observed class (in $\%$) are marked. On the y-axis, the accuracy of the observed class is displayed in terms of F-measure.

in Figure 5, only with a different sorting, as just described. This figure reveals the following aspects:

1) The general tendency to be affected by chance decreases with increasing number of classes (r). This has been discussed in Section 5.1.
2) Within a particular number of classes, the accuracies vary in considerable ranges, which depend on the number of classes.
3) Datasets with small number of classes seem to be more affected by the class size than datasets with large number of classes, which is discussed in more detail in the next paragraph.

Figure 6 (B) and (C) show details of the first segment of (A), i.e. 2 classes, and the last segment of (A), i.e. 9 classes, respectively. While the accuracies in (B) rapidly increase with increasing class size, they show a slow increase in (C). However, a careful look at the domain of each of the graphs explain the observation. The class sizes in (A) reach up to 95%, while the maximum class size in (C) is about 61%. This is because it is less probable for a class to dominate a dataset consisting of many classes than one consisting of a lower number of classes. In other words, for a given class q to dominate a data set of r classes, there should be $r-1$ classes with significantly smaller class size than of q. This means that the more classes a dataset has, the less probable that one class dominates the it, which means in turn that the class imbalance tends to be less probable.

5.3 Impact of the columns-rows ratio ρ

Figure 6 still ignores some elements that govern the variation in the accuracy for a fixed number of classes. Therefore, in the remainder of this section, we will exclude the impact of class size (class imbalance) to be able to observe the impact of the remaining factors.

To exclude the impact of class imbalance, we generated a new group of $1,000$ random datasets according to Definition 1, with the restriction that all datasets have equal class sizes, i.e. *balanced* classes. We call this group BALANCED. We trained classification models based on these datasets and observed their accuracies in terms of F-measure. Each point in Figure 7 corresponds to the score of one model, where the models are sorted first according to ρ (the ratio between the number of columns m and the number of rows n) and then according to the number of classes r. The points are divided into segments, in which the models trained on a certain number of classes are grouped. Within each segment one can observe that F-measure values vary and there is a tendency of increasing F-measure with increasing the ratio ρ. This variation has clearly nothing to do with class size, being it equal for all classes. One can also observe that the influence of ρ on the accuracy decays with increasing r. It is almost negligible in the last segment, i.e. for $r=9$ classes.

It is remarkable that increasing the number of classes (r) leads to a decreasing influence of both the ratio ρ and also the class size (Q), as shown in Section 5.2.

5.4 Testing with genomics data

The aim of this section is to ensure that the observations described as well as the estimators presented in Sec. 4 are not just a bias of the particular distributions used to generate the random datasets, i.e. the normal and uniform distributions.

For this, we generated shuffled real datasets (RDREAL) according to Definition 3 as follows: 200 datasets were generated by modifying a real microarry gene expressions with $\sim 16k$ genes (features) and contains 80 samples from patients who died as a result of neuroblastoma cancer after different survival times. The data has been divided into two classes (long and short event free survival) as a binary classification task. After that, the binary target class of the dataset has been shuffled according to Definition 3 such that the dataset became meaningless regarding the target.

Figure 8 is box and whisker plot that shows the accuracies of models trained on the datasets of the group RDREAL. The figure also shows the predicted accuracy based on the dataset dimensionality using Eq. (4). The predicted value deviates from the mean of the actual accuracies by about 5%, which indicates that Eq. (4) extends its applicability and accuracy to real-world distributions.

5.5 Chance influence on Regression

The aim of this section is first to demonstrate the high impact of chance correlation on regression models and second to motivate extending the formal analysis to regression in a future work. In particular, we empirically show that the accuracy (measured as root mean squared error) of regression models trained on random data (uniform and Gaussian) can be estimated using the pessimistic chance estimator, Eq.(6), considering a two class classification.

To this end, we generated two random datasets with numeric targets. RDUNIF consists of 1000 datasets generated according to Definition 2, where the number of features and samples vary from 10 to 1000. The targets are random numbers. Feature values and target values are drawn from

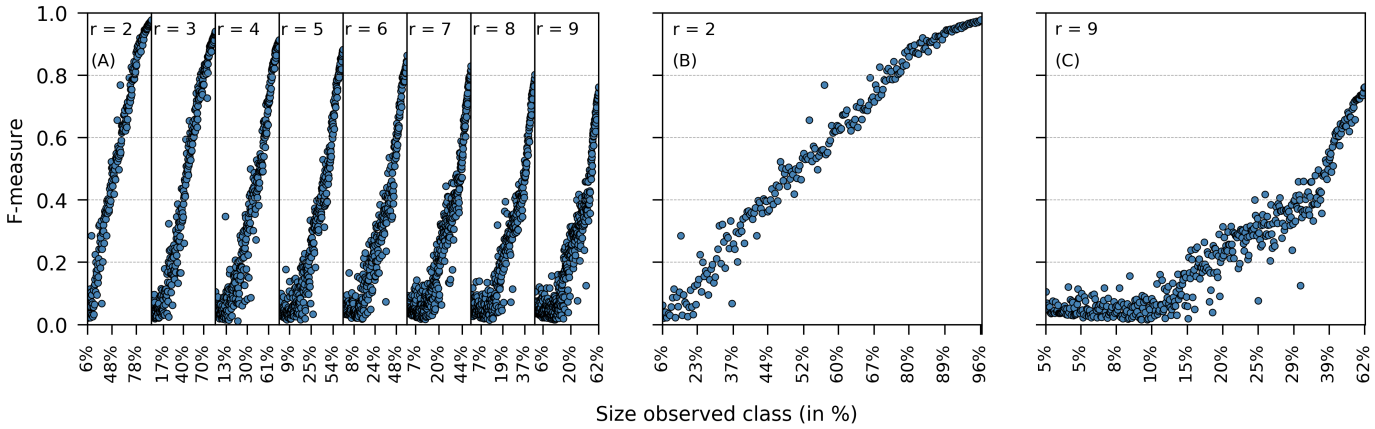

Fig. 6. Relation between accuracy and class size while sorting according to number of classes. (A) The setting is as in Fig. 5, except that measurements are sorted according to number of classes after sorting them according class size. (B) Detailed plot for the first segment, i.e. two classes. (C) Detailed plot for the last segment, i.e. 9 classes.

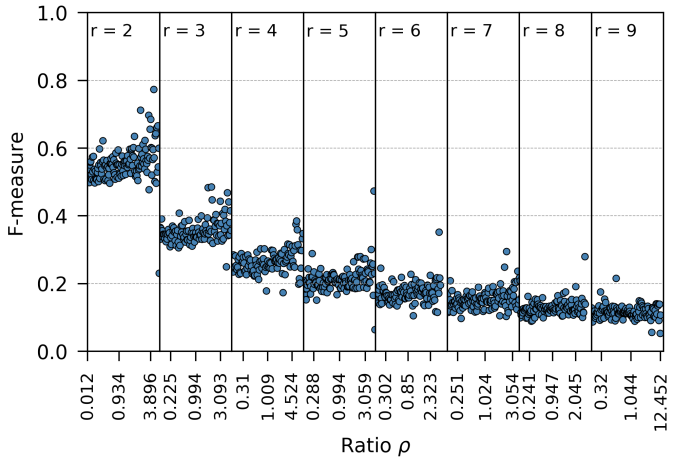

Fig. 7. Classification performance in terms of F-score for models trained on datasets of the BALANCED group. Data points are first sorted according to ρ (the ratio between the number of columns m and the number of rows n) and then according to the number of classes r.

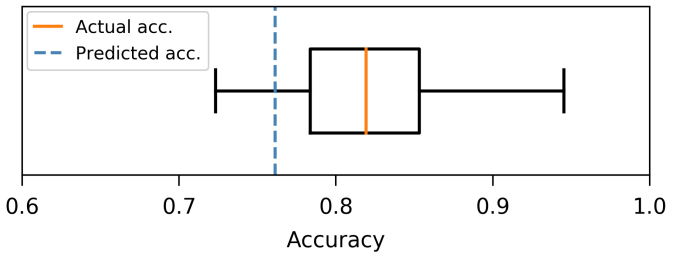

Fig. 8. Box and whisker plot showing the accuracy of prediction models trained on RDREAL. The boxes extend from the lower to upper quartile values, while the whiskers indicate the data range.

a uniform distribution in the interval $[0, 1]$. RDGAUSS consists of 1000 datasets, as in RDUNIF, but with target values drawn from a Gaussian distribution with variance $\sigma = 1$ and mean $\mu = 0$. The predicted values are evaluated against the values of the target using the mean squared error (MSE), to find the accuracy of the model obtained purely by chance.

Figure 9, yellow and red points show the accuracies (RMSE values) of regression models, trained on the datasets of the groups RDUNIF and RDGAUSS respectively. In the first group (yellow), each point represents the RMSE of a model trained on one of the 1000 datasets from the RDUNIF and in the second group (red), each point corresponds to a model trained on one of the 1000 datasets of the RNDGAUSS.

A deeper look at Figure 9 shows that the RMSE values of the RDUNIF have a range, approximately, between 0 and 0.3, while the values of the RDGAUSS group range between 0 and 1. An RMSE of zero means a perfect fit, it corresponds to models providing a perfect prediction. In the case of the present experiment, these are the models that are mostly affected by chance. On the other side, the values of about 0.3 (yellow) and 1.0 (red) represent the models with least accuracy, which in our case corresponds to the ones that are least affected by chance. But where do these values (i.e. about 0.3 and 1) stem from? To answer this question, let us consider the distributions of the target values in each group and the definition of the RMSE. The target values in RDUNIF group have a uniform distribution in $[0, 1]$. A uniform distribution in $[a, b]$ has a standard deviation $\sigma = \sqrt{(b-a)/12}$. In our case we get $\sqrt{1/12} \approx 0.289$. When the prediction algorithm performs pure randomly, the RMSE between the uniformly distributed values and the predictions is also uniformly distributed and has an expectation value equal to the standard deviation of the values. The same holds for the RDGAUSS group, but with an expectation value of 1, because the target values are distributed normally with $\sigma = 1$ as it was configured in the data generation.

Now, let us come back to the aim of this section. Given a random dataset with real target according to Definition 2 of the shape $m \times n$, we want to test whether we can estimate the accuracy of regression models trained on this dataset in terms of its dimensions m and n using the pessimistic chance estimator for classification, presented in Section 4.2 (Eq.(6)). The basic idea is to transfer the chance estimation method for classification to the regression task in the two steps:

I. Considering the chance estimate of classification with two classes (i.e. a classification of dimensionality $m \times$

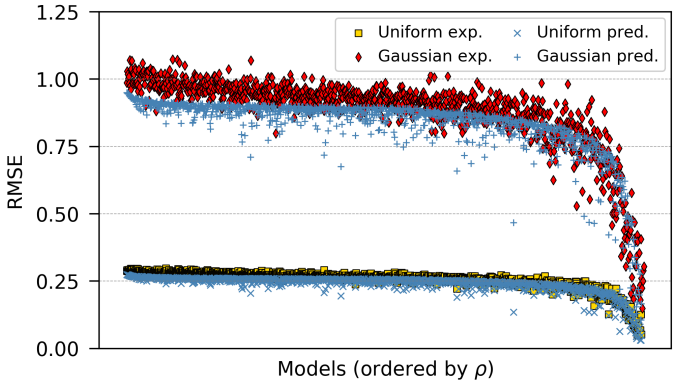

Fig. 9. Actual root-mean-square error (RMSE) values for regression experiments on the RDUNIF (yellow squares) and RDGAUSS (red diamonds) datasets together with the predicted RMSE values by the estimator Eq. (10) (blue plus and cross symbols). Actual and predicted values reveal a Pearson correlation coefficient of 0.92 and 0.913 for the two groups respectively.

$n \times 2$) and equal class sizes, i.e. by substituting $p = 1/2$ in Eq. (6). This results in

$$\breve{\phi} = 1 - \frac{1}{n}\sum_{k=1}^{n}\left[\sum_{t=0}^{k-1}\binom{n}{t}\frac{1}{2^n}\right]^m. \tag{8}$$

II. Normalizing this estimate to fit in the range of RMSE as described above. First, since RMSE is a distance measure that increases with decreasing model accuracy, in contrast to Eq.(6) that is proportional to the accuracy, the normalization should perform a reversion. Second, the normalization should bring the range of Eq.(6), which is $[1/2, 1]$, to the range of RMSE described above, i.e. $[0, \sigma]$.

Putting I. and II. together, the normalization that converts $\breve{\phi}$ (pessimistic estimate for classification) to ρ (estimate for regression accuracy measured by RMSE), is given by:

$$\rho = 2\sigma(1 - \breve{\phi}). \tag{9}$$

which directly leads to

$$\rho = 2\sigma\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{t=0}^{k-1}\binom{n}{t}\frac{1}{2^n}\right]^m \tag{10}$$

Figure 9 shows the RMSE values of the regression models of the groups RDUNIF and RDGAUSS, together with the values predicted using Eq. (10) (blue points). The overall Pearson correlation coefficient between the experimental RMSE values and their predictions is about 0.92. This motivate to formally extend the analysis in Sec. 4.2 to the regression task in a future work.

5.6 Normal Approximation

The estimators in Equations (5), (6), and (7) are functions of the binomial distribution, which can be approximated by a normal distribution $\mathcal{N}(\mu, \sigma)$ with mean $\mu = np$ and variance $\sigma = np(1-p)$, given that n is large enough in relation to p and $(1-p)$. In general, the sample size is considered large enough if: $np \gtrsim 10$ and $n(1-p) \gtrsim 10$, which is often the case for real-life datasets. This means that the equations above can be approximated by replacing the binomial part as follows

$$\binom{n}{x}p^x(1-p)^{n-x} \rightarrow \frac{1}{\sqrt{2\sigma\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{11}$$

where p is the representative single probability, $\mu = np$ and $\sigma = np(1-p)$. This form of the estimators is significantly simpler in terms of calculation time.

5.7 Correction for chance

Evaluation metrics that correct for chance are not new in this field. For instance, the Cohen's Kappa metric [47] calculates the agreement between two raters, thereby taking into consideration the expected agreement by chance:

$$\text{Kappa} = \frac{A_0 - A_e}{1 - A_e} \tag{12}$$

where A_0 is the measured agreement between two raters and A_e is the hypothetical probability of chance agreement. For a classification with n objects and r classes, A_e is defined as:

$$A_e = \frac{1}{n^2}\sum_{i=1}^{r} N_{1i}N_{2i}, \tag{13}$$

where N_{ji} the number of objects predicted by rater j as belonging to class i. As an example, assume we have n objects, where half of them belong to class A and the other half to class B. A naive classifier that assigns all objects to class A would have an accuracy of 0.5, but the Kappa measure will be 0 because the expected true guessing (0.5 for two classes) is subtracted from the score.

Equation (13) shows that Kappa considers only the number of the classes and the class distribution to estimate the chance, which means that other factors, in particular the number of features in the dataset, is not considered. In other words, Kappa calculates the hypothetical probability of chance based only on the object-class distribution. To demonstrate this fact, we evaluated the classification performance on the datasets of the RDCAT group also using the Kappa measure and plotted them beside the F-measure in Figure 10. It is clear how the F-measure values have been shifted down by the Kappa measure, but these values are still not zero. The remaining classification performance, which has not been corrected by it, stems from the factor not considered by Kappa: the shape of the dataset, i.e. the ratio ρ.

We propose in this work estimators that go beyond the well-known Cohen's Kappa metric by considering elements of a dataset that are not taken into account by the latter, namely its number of samples and features. This is an important refinement for datasets that have significantly more features than samples. Along the lines of this metric, we propose to subtract the accuracy by chance (estimator value) from the accuracy measure. Assume S to be the performance score of a model measured using some evaluation metric that does consider for chance correction, like accuracy or F-measure. Then, the corrected score S_c would be

$$S_c = S - \phi.$$

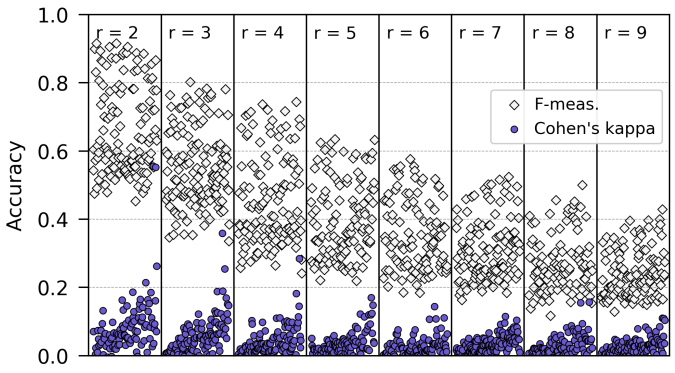

Fig. 10. Cohen's Kappa (purple circles) and F-measure (white diamonds) as evaluation of prediction models trained on the RDCAT datasets. The kappa measure corrects the values for chance by shifting them towards zero, but does not correct the accuracy increase by chance stemming from ρ.

Note that in order to use ϕ for score correction, the original score should be normalized to the range $[0, 1]$, if it is not already in this range. Also note that the S_c could be negative, which denotes the case of "worse than random" (analog to the negative correlation).

S and S_c can be used to describe the model confidence in the best and worst case, respectively, effectively describing an accuracy range based on the observed score and the dataset parameters.

5.8 Guidelines for chance mitigation and interpretation of model accuracy

Based on the formal analysis of chance as well as the observations and empirical results, we provide general guidelines for mitigation of chance influence and right interpretation of evaluation metrics to avoid misleading high accuracies when the data is likely to contain considerable amount of chance.

1) **Estimate the chance:** Consider the dimensions of the dataset at experiment design to estimate the extent of chance in the dataset using the proposed estimators to decide which data to use and how many instances you need to achieve the required level of confidence.
2) **Report the chance values:** We encourage researchers not only to estimate the impact of chance, but also to consider it when interpreting their results, and report it in their works.
3) **Consider correction for chance:** If the data is imbalanced but low dimensional, use Kappa measure to evaluate the model. If the data is high dimensional, Kappa is not sufficient to correct for chance. In this case use the proposed chance estimators to correct for chance as described in Section 5.7
4) **Increase the number of classes:** Section 5.1 shows the strong influence of the number of classes on chance extent. The more classes in the target, the less chance in the data. Where possible, try to re-aggregate/prepossess your data such that it has as much classes as possible.
5) **Encode numeric target to multi-class:** For the same reason in 4) it is recommended, where possible, to encode numerical target to multi-class target by using kind of interval coding. This obviously means converting a regression to a classification task.

5.9 Impact of the proposed analysis

The formal analysis and the chance estimators provide a foundation to quantify the value of a dataset as well as the confidence of models created based on it. We believe that the proposed work is relevant in the following aspects.

- **Actual accuracy as opposed to model ranking**. We believe it is very valuable to know the extent of *real* accuracy in order to know how much to rely on the obtained results (e.g. in the medical sector). In other words, the proposed methods aim to go beyond algorithms comparison and consider the performance of a specific algorithm—regardless of its rank among other algorithms—to determine its reliability. This is of obvious help in case of scientific publications, when discerning the validity of claims based on results obtained with a given dataset, or in fields such as the medical one, where decisions can have serious consequences. In contrast to existing measures that correct for chance, like the Kappa measure (cf. Sec. 5.7), our proposed chance estimators go beyond class distribution and consider also the shape of the dataset in terms of number of features and number of instances.
- **Generalisation.** A key opportunity for the application of our results is in situations where researchers (and practitioners) build models achieving high performance and they need to truthfully interpret these results, so that they do not over-generalise what they have found. Also, our results support the recent attention that has been given to the irreproducibility problem affecting different areas of science (see e.g. Sec. 2.3 "Critical Research" and references therein). In this context, our estimators can be also used to verify the reliability of published research results. This can be achieved by applying the chance estimators on the parameters of the dataset based on which results have been concluded.
- **Experimental design.** In the experimental design phase, our proposed chance estimators can help to understand and estimate how many instances are needed to reach a certain degree of confidence in the results.
- **Data valuation.** In the context of databases and data markets, an active research field deals with the challenging question of assigning (monetary) value to datasets. This depends on several factors like size, timeliness, and completeness, to name a few. The proposed chance estimators provide another formal property for data value assessment, namely the extent of chance in a dataset.

CONCLUSION

We empirically showed that experiments conducted on wide datasets can frequently be susceptible to a significant impact of chance influence. If ignored, chance can be a key factor leading to spuriously accurate but not generalisable models. We proposed estimators quantifying the extent of chance in a categorical dataset based on its parameters, namely (i) the dimensions, i.e. number of features and

samples, (ii) the number of classes, and (iii) the class distribution. The results of our experiments indicate high accuracy of these estimators, with the most accurate estimator reaching a Pearson correlation coefficient of 0.979 (95% CI $[0.976, 0.981]$).

We presented guidelines for mitigating the influence of chance by researchers working with wide datasets. These guidelines ask researchers to estimate chance, consider it when interpreting their results, and report it in their works.

ACKNOWLEDGMENTS

This work was carried out within the Austrian Research Promotion Agency (FFG) COIN Networks project VISIOMICS and the Horizon 2020 Safe-DEED project (GA No. 825225).

REFERENCES

[1] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proceedings of the National Academy of Sciences*, vol. 103, no. 15, pp. 5923–5928, 2006. [Online]. Available: http://www.pnas.org/content/103/15/5923

[2] X. Fan, L. Shi, H. Fang, Y. Cheng, R. Perkins, and W. Tong, "Dna microarrays are predictive of cancer prognosis: A re-evaluation," *Clinical Cancer Research*, vol. 16, no. 2, pp. 629–636, 2010. [Online]. Available: http://clincancerres.aacrjournals.org/content/16/2/629

[3] J. P. A. Ioannidis, "Why most published research findings are false," *PLOS Medicine*, vol. 2, no. 8, 08 2005. [Online]. Available: https://doi.org/10.1371/journal.pmed.0020124

[4] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488 – 492, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140673605178660

[5] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, "Optimal number of features as a function of sample size for various classification rules," *Bioinformatics*, vol. 21, no. 8, pp. 1509–1515, 2005. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bti171

[6] X. Lian and L. Chen, "General cost models for evaluating dimensionality reduction in high-dimensional spaces," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 10, pp. 1447–1460, Oct 2009.

[7] H. Kim, B. S. Choi, and M. Y. Huh, "Booster in high dimensional data classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 29–40, Jan 2016.

[8] R. Weber, H.-J. Schek, and S. Blott, "A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces," in *Proceedings of the 24rd International Conference on Very Large Data Bases*, ser. VLDB '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, p. 194205.

[9] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *Database Theory — ICDT'99*, C. Beeri and P. Buneman, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 217–235.

[10] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Database Theory — ICDT 2001*, J. Van den Bussche and V. Vianu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 420–434.

[11] C. Hsu and M. Chen, "On the design and applicability of distance functions in high-dimensional data space," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 4, pp. 523–536, 2009.

[12] R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton university press, 2015, vol. 2045.

[13] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nature reviews cancer*, vol. 8, no. 1, pp. 37–49, 2008.

[14] Y. Wang, D. J. Miller, and R. Clarke, "Approaches to working in high-dimensional data spaces: gene expression microarrays," *British journal of cancer*, vol. 98, no. 6, pp. 1023–1028, 2008.

[15] A. Sinha, G. Hripcsak, and M. Markatou, "Large datasets in biomedicine: a discussion of salient analytic issues," *Journal of the American Medical Informatics Association*, vol. 16, no. 6, pp. 759–767, 2009.

[16] J. Kuligowski, D. Perez-Guaita, J. Escobar, M. Guardia, M. Vento, A. Ferrer, and G. Quintas, "Evaluation of the effect of chance correlations on variable selection using partial least squares-discriminant analysis," vol. 116, pp. 835–40, 11 2013.

[17] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6562–6566, 2002. [Online]. Available: http://www.pnas.org/content/99/10/6562

[18] J. Hua, W. D. Tembe, and E. R. Dougherty, "Performance of feature-selection methods in the classification of high-dimension data," *Pattern Recognition*, vol. 42, no. 3, pp. 409 – 424, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320308003142

[19] K. K. Dobbin, Y. Zhao, and R. M. Simon, "How large a training set is needed to develop a classifier for microarray data?" *Clinical Cancer Research*, vol. 14, no. 1, pp. 108–114, 2008. [Online]. Available: http://clincancerres.aacrjournals.org/content/14/1/108

[20] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. R. Golub, and J. P. Mesirov, "Estimating dataset size requirements for classifying dna microarray data," *Journal of Computational Biology*, vol. 10, pp. 119–142, 2003.

[21] R. Fano, *Transmission of Information: A Statistical Theory of Communications*. Cambridge, MA: The MIT Press, 1961.

[22] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[23] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, ser. STOC '98. New York, NY, USA: Association for Computing Machinery, 1998, p. 604613. [Online]. Available: https://doi.org/10.1145/276698.276876

[24] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 5, no. 5, pp. 363–387, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11161

[25] D. Francois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 7, pp. 873–886, 2007.

[26] F. Angiulli and C. Pizzuti, "Outlier mining in large high-dimensional data sets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 203–215, 2005.

[27] W. Li, Y. Zhang, Y. Sun, W. Wang, M. Li, W. Zhang, and X. Lin, "Approximate nearest neighbor search on high dimensional data experiments, analyses, and improvement," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1475–1488, 2020.

[28] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 12, pp. 1624–1637, 2005.

[29] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37 – 52, 1987, proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0169743987800849

[30] H. H. Harman, *Modern factor analysis*. University of Chicago press, 1976.

[31] Huan Liu and Lei Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[32] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 1–14, 2013.

[33] Huan Liu and R. Setiono, "Feature selection via discretization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 9, no. 4, pp. 642–645, 1997.

[34] M. A. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," 2000.

[35] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
[36] Tin Kam Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, 1995, pp. 278–282 vol.1.
[37] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance," *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, p. 8, Feb 2012. [Online]. Available: https://doi.org/10.1186/1472-6947-12-8
[38] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft, and J. Popp, "Sample size planning for classification models," *Analytica Chimica Acta*, vol. 760, pp. 25 – 33, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003267012016479
[39] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and information systems*, vol. 12, no. 1, pp. 95–116, 2007.
[40] M. Clark and R. Cramer, "The probability of chance correlation using partial least squares (pls)," *Molecular Informatics*, vol. 12, no. 2, pp. 137–145, Jan. 1993.
[41] A. A. Taha, A. Bampoulidis, and M. Lupu, "Chance influence in datasets with a large number of features," in *Data Science – Analytics and Applications*, P. Haber, T. Lampoltshammer, and M. Mayr, Eds. Wiesbaden: Springer Fachmedien Wiesbaden, 2019, pp. 21–26.
[42] P. Kraft, E. Zeggini, and J. P. A. Ioannidis, "Replication in genome-wide association studies," *Statist. Sci.*, vol. 24, no. 4, pp. 561–573, 11 2009. [Online]. Available: https://doi.org/10.1214/09-STS290
[43] A. Casadevall and F. C. Fang, "Reforming science: Methodological and cultural reforms," *Infection and Immunity*, vol. 80, no. 3, pp. 891–896, 2012. [Online]. Available: https://iai.asm.org/content/80/3/891
[44] B. A. Nosek, J. R. Spies, and M. Motyl, "Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability," *Perspectives on Psychological Science*, vol. 7, no. 6, pp. 615–631, 2012, pMID: 26168121. [Online]. Available: https://doi.org/10.1177/1745691612459058
[45] D. Colquhoun, "An investigation of the false discovery rate and the misinterpretation of p-values," *Royal Society Open Science*, vol. 1, no. 3, p. 140216, 2014.
[46] ——, "The reproducibility of research and the misinterpretation of p-values," *Royal Society Open Science*, vol. 4, no. 12, p. 171085, 2017.
[47] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, p. 37, 1960.

Abdel Aziz Taha is a data scientist at Research Studios Austria. He received his PhD in Data Science with distinction from the Vienna University of Technology in 2015. The dissertation of Abdel Aziz Taha offers solutions for problems in the analysis of large data under extreme conditions, such as handling metric distortions and enabling highly efficient computation of complex data as well as data with very high dimensionalities . Abdel Aziz Taha has been involved in research and development projects in the field of machine learning, including genomics, medical diagnostics, fraud detection, industrial automation, anomaly detection and prediction, and large data marketing.

Luca Papariello received the BSc and MSc degrees in physics from ETH Zurich, Switzerland. In the same university, he received the PhD degree in physics from the Institute for Theoretical Physics in 2018. He is currently researcher and data scientist in the Research Studio Data Science of the RSA FG, where he is working on national and European projects with focus on machine learning, natural language processing, and data mining in general.

Alexandros Bampoulidis received his BSc degree in informatics from the Alexander Technological Education Institute (ATEI) of Thessaloniki, Greece (currently known as International Hellenic University) in 2014. He is currently working toward his MSc degree in business informatics from TU Wien in Vienna, Austria. His research interests include (big) data science, privacy, information retrieval.

Petr Knoth is a researcher leading R&D teams working in the domains of text-mining, data science, digital libraries and open access/science. Petr has a deep interest in the use of AI to improve research workflows. Petr is currently a Senior Research Fellow at The Open University, UK and Director Data Science at Research Studios Austria. In 2011, Petr founded and since then has been the Head of CORE (core.ac.uk), the worlds largest full text aggregation of open access papers making them available for people to access and machines to text-mine. Petr authored or co-authored over 70 peer-reviewed publications in international venues and received two best paper awards at top conferences in his field. Previously, he worked as a Senior Data Scientist at Mendeley on information extraction and content recommendation for research.

Mihai Lupu is, since January 2018, the Studio Director of the Data Science Studio at Research Studios Austria Forschungsgesellschaft. Before that he has been a researcher at the TUWien as well as a private entrepreneur, consulting small and large companies on search technology, with focus on search in the intellectual property domain. He graduated from the Singapore-MIT Alliance in 2008 and since then has published over 100 publications and three books on search technology. He is now the co-coordinator of Data Market Austria and the Coordinator of the H2020 Safe-DEED project.